Reliability Analysis on Measurement Data from Scientific Clusters

Project Description

Technical advances in scientific clusters have caused increasing scales of data accesses and movement, number of nodes involved in job executions, and exploited parallelism in applications on many cores. These changes lead to unforeseen increased scales of interactions and communications in software executions between hardware resources and nodes in a cluster. However, technical advances have not reached the reliability of hardware components and softwares in the same scale. Therefore, the failure rates have been dramatically increased due to the combinations of the increased scales of software and hardware executions and the lack of improvements in their reliability in the same scale.

Despite these increased failures, system developers and administrators are overwhelmed by large-scale logs to take actions for reducing failures and their impacts such as wasted resources due to failures. It is challenging to analyze the failures on large scientific clusters due to the size of logs and noises in their measurements from the interactions and interferences in the job executions. Since a node can be assigned multiple tasks from different jobs, the measured executions can be noisy because of performance interferences in shared hardware resources from co-located tasks from multiple jobs. This indeterministic variances and noises due to interferences make failure prediction more challenging.

The goal of this project is analyzing failure cases from the measurement data. The analysis can identify common patterns of failures and historical characteristic patterns of executions from failures. These identified patterns can be compared to the ongoing execution characteristics to predict failures in online fashion. This can improve the reliability of scientific clusters and application executions.

Task Goals

The main research goal is to analyze failures from the measurement data of the NERSC scientific cluster and identify patterns and their associations to improve reliability. It includes following items:

- Analysis of failure cases from the collected measurement data
- Identification of patterns in failure cases
- Analysis of identified patterns to identify possible associations between the patterns

Task Requirements

- Proficient in a programming language, such as python
- Good problem solving skills and communication skills
- Enthusiastic about exploring new ideas and identifying research challenges

About the group

The Scientific Data Management (SDM) group at Lawrence Berkeley National Laboratory develops technologies and tools for efficient data access, data storage, data analysis, and management of massive scientific data sets. We are currently developing storage resource management tools, data querying technologies, in situ feature extraction algorithms, data analysis algorithms, along with software platforms for exascale data. The group also works closely with application scientists to address their data processing challenges. These tools and application development activities are backed by active research efforts on novel algorithms for emerging hardware platforms.